**一．本周进行的主要工作内容**

1.针对毕设题目《面向特定商品的用户购买原因和用途识别》，分析题目，初步确定了思路为结合自然语言处理（NLP）的相关知识原理对特定商品的用户评论进行文本分析并分类，实现对用户购买原因和用途的识别。

    2.遴选了适宜的特定商品。为使数据特征更加鲜明、更富有多样性，初步计划选择三种特定商品，并对每种特定商品分别进行数据挖掘和文本分析。

|  |  |  |
| --- | --- | --- |
| 商品名称 | 是否具有主观预测性 | 用途和购买原因多样性 |
| 微波炉 | 有，与做饭有关 | 较为多样 |
| 笔记本电脑 | 无 | 较为多样 |
| 化妆品 | 有，与护肤美容有关 | 较为单一 |

可以看出，这三种商品特征鲜明，且相互之间各有相似点与不同点，这有利于使得本毕设的最终分析结果囊括性更加丰富和全面。

3.初步确定了要使用的技术手段，有：

（1. 爬虫技术用于数据获取

（2. 利用Python-jieba进行数据初步处理，如清洗掉非中文字符、文本切片、中文词性标注等。

（4.利用Python-pandas进行文本词频统计、数据导出等

（3. 利用LDA模型进行主题分析

4.学习了相关理论知识。

（1. 学习了爬虫技术。

（2. 学习了利用Python-jieba的切片、词性标注技术。

（3. 学习了利用Python-pandas的词频统计、数据导出技术。

（4初步了解了LDA模型。

学习来源：知网相关论文、Blibli相关公开课程、Python官方网站的pypi

5.根据理论知识完成了既定计划进度。利用爬虫技术从购物网站对特定商品的用户评论进行了数据爬取，并进行了初步数据清理，对得到的文本利用jieba进行了文本词语的切片以及词性的标注，并利用pandas完成了各特定商品的词频统计及本地保存。

**二．下周计划**

继续学习LDA，完善并改进文本的分析过程及处理方式。